APPENDIX F

Evidence for the Reliability and Validity of Scores from the Washington Assessment of Student Learning (WASL)

Presented to:

Washington Roundtable
December 11, 2002

Catherine S. Taylor
University of Washington

Reliability and Validity of Scores: Common Sense and Psychometrics

Dictionary Definitions:

Valid: logical or empirical truth

Reliable: dependable

Logical Truth

Can we make an argument that this score truly is a measure of the knowledge, skills, and strategies we wanted to measure?

1. Building tests requires careful thinking about 'the game to be played'. The test developer must ask her/himself:

- "What are the important skills, strategies, and concepts that define the subject to be tested?"
- "What are the most appropriate ways to assess the knowledge, skills, and strategies?"

Empirical Truth

Can we show evidence that this this score truly is a measure of the knowledge, skills, and strategies we wanted to measure?

1. To gather empirical evidence, the test developer must ask him/herself:

- "Do the scores mean what I think they mean? Would students perform the same way on other tests that measure the same or similar content?"
- "Is there another possible explanation for scores? Is some phenomenon, other than knowledge, skills, and strategies in the subject area, affecting performance on the test?"

Dependability

Can we show evidence that if I gave this or a similar test again, the student would get about the same score?

POTENTIAL SOURCES OF EVIDENCE FOR THE VALIDITY OF TEST SCORES

There are several ways that test developers get *evidence for validity* of test *scores*:

1. Content Validity Evidence. Professional judgment about whether the content measured is appropriate and represents the range of what examinees should know and be able to do

2. Construct Validity[1] Evidence. Strong correlations (.70-.80) between scores from different tests that are supposed to be measuring the same or about the same content and skills[2]

3. Construct Validity Evidence. Moderately strong correlations (.50-.70) between scores from different tests that measure related content and skills (for example, reading comprehension and listening comprehension)

4. Construct Validity Evidence. Moderate correlations (.40-.60) between scores on tests intended to measure very different content and skills (writing and mathematics)

5. Construct Validity Evidence. Mathematical analysis of the *patterns of examinee performance* to see whether the abilities that 'cause' performance on tests measuring the same or similar skills result in similar performance from examinees (also called 'factor' analysis).

[1] A *construct* is the definition we give to what we are testing – e.g., reading comprehension, mathematics problem solving, mathematics computation, writing skills. The construct validity question is, "Do we have sufficient evidence to believe that *scores* from this test really tell us whether students can comprehend what they read, can solve mathematics problems, understand scientific concepts, etc.?"

[2] NOTE: In evaluating the numbers in the tables, keep the following in mind:

- Correlations can range from –1.00 to 1.00.
- A correlation of –1.00 indicates that total scores from one test reflect the opposite of total scores on a second test.
- A correlation of 1.00 indicates the total scores on two tests are the same or very nearly the same.

CONTENT VALIDITY EVIDENCE

1. Accepted strategies for obtaining professional judgment about whether the content measured is *appropriate* and *represents the range* of what examinees should know and be able to do.

   - Check the item specifications to make certain they match the content standards (EALRs)
   - Check test specifications to make certain that they represent the range of knowledge and skills described in the content standards
   - Check *items* and *scoring rules* (rubrics) to make certain items actually match the item specifications and measure the content standards; to make certain the scoring rules match the content standards.

2. All of these steps were taken in the multiple stages of development of WASL.
3. Reviewers were teachers guided by professional testing specialists
4. Additional external evaluation studies that have been done[3]:
   - evaluation of grade 10 mathematics WASL
   - evaluations of the reading, mathematics, and science EALRs
5. Additional evaluation to be done:
   - evaluations of all EALRs, frameworks against standards for National Assessment of Educational Progress and objectives for Iowa Test of Basic Skills/Iowa Test of Educational Development.
6. One step where WA state went beyond many states is in the public scrutiny of the state standards (EALRs) and revisions based on widespread public input

[3] Standard practice in test development

CONSTRUCT VALIDITY EVIDENCE

Examine correlations between scores from different tests that are supposed to be measuring the same or *about the same* content and skills

- If two tests are supposed to measure exactly the same content and skills (for example, two forms of the Iowa Test of Educational Development [ITED]), the correlations should be *very high* (about .90)
- If two tests are supposed to measure similar knowledge and skills but also have differences in terms of the targeted knowledge and skills, the correlations should be *strong* but *not too high* (between .70 and .80)

Scores from Grade 10 WASL Correlated with Norm-Referenced Test Scores

WASL Year	Content	Test Pair	Correlation
2001	Math	ITED[4] (spring grade 9) and WASL (spring grade 10)	.796
2001	Reading	ITED (spring grade 9) and WASL (spring grade 10)	.744

[4] Iowa Test of Educational Development

CONSTRUCT VALIDITY EVIDENCE

Examine correlations among scores from tests that are supposed to be measuring *related* knowledge and skills (desirable correlations = .50 - .70)

Scores from Grade 10 WASL Correlated with Related WASL Tests

WASL Year	Test(s)	Correlation
1999	WASL Reading and WASL Listening	.649
1999	WASL Reading and WASL Writing	.646
2000	WASL Reading and WASL Listening	.652
2000	WASL Reading and WASL Writing	.693
2001	WASL Reading and WASL Listening	.634
2001	WASL Reading and WASL Writing	**.725**

CONSTRUCT VALIDITY EVIDENCE

Examine correlations among scores from tests that are supposed to be measuring *very different* knowledge and skills (expected correlations = .40 - .60)

Scores from Grade 10 WASL and Norm Reference Tests

WASL Year	Test(s)	Correlation
2001	WASL Listening and WASL Writing (2001)	.536
2001	WASL Math and WASL Writing (2001)	*.648*
2001	WASL Math and WASL Reading (2001)	**.733**
2001	WASL Math and ITED Reading (2001)	*.692*
	ITED Math and ITED Reading (2000)	**.741**

These results show:

A stronger than expected relationship between reading and mathematics scores within ITED, within WASL, and, sometimes, between WASL and ITED

A stronger than expected relationship between WASL mathematics scores and WASL writing

This required follow-up studies to investigate **potential** explanations for results:

a) ITED and WASL demand reading (of words and numbers) in the mathematics tests

b) WASL mathematics demands the skills that are demanded in a writing assessment

CONSTRUCT VALIDITY EVIDENCE

Mathematical analysis of the *patterns of examinee performance* (also called 'factor' analysis).

I. Each year strand scores for WASL reading and mathematics are analyzed using factor analysis

   Consistent results each year:

   1. Mathematics factor composed of scores from:
      a) Number sense
      b) Measurement
      c) Geometric sense
      d) Probability and statistics
      e) Algebraic sense
      f) Mathematical problem solving
      g) Mathematical reasoning
      h) Mathematical communication
   2. Reading factor composed of scores from:
      a) Main ideas and details of fiction
      b) Analysis and interpretation of fiction
      c) Critical thinking about fiction
      d) Main ideas and details of nonfiction
      e) Analysis and interpretation of nonfiction
      f) Critical thinking about nonfiction
   3. Writing factor composed of scores from:
      a) Content, organization and style in writing
      b) Writing Conventions

CONSTRUCT VALIDITY EVIDENCE

II. During 2001, one Grade 10 study was conducted looking at patterns of performance on subscores for ITED and WASL:

The analysis showed two underlying factors

a) Language arts factor (WASL reading, listening, and writing strand scores, ITBS Literary and Vocabulary subtest scores required the same underlying knowledge and skills)

b) Mathematics factor (WASL mathematics strand scores, ITBS Mathematics Quantitative subtest scores required the same underlying knowledge and skills)

The study provides evidence that supports the claim that Grade 10 WASL mathematics test measures mathematics and Grade 10 WASL reading, writing, and listening tests measure the language arts. While reading is needed in mathematics, it is not a reading test.

ADDITIONAL CONTENT VALIDITY STUDIES[5]

1. Study to review of the reading, mathematics and science EALRs (study being conducted by MCREL)

2. Study to examine match between WASL mathematics items and EALRs (study conducted by SRI)

3. Study comparing the content of WASL mathematics and reading assessments with community college placement exams (study conducted by State Board of Community Colleges)

4. Study to examine the 'drift' of scaled scores over time (being conducted by UW)

Early results suggest that the score scale is *extremely* stable over time; students would earn the SAME scale score regardless of how scaling is done (i.e., a 400 is a 400 is a 400 regardless which year the test is administered)

5. Studies to examine validity of the strand scores (being conducted by UW)

Early results suggest that scores could also be presented based on the thinking skills involved in mathematics (e.g., recall of simple rules, solving complex multi-step problems) or based on type of text read (i.e., informational vs. narrative)

6. Study to examine whether traits other than mathematics affect mathematics scores (studies conducted by UW):

   - Reading study shows that students with reading difficulties tend get their scores from items that have visual displays; students without reading difficulties get their scores from items that have verbal text; test is a balance of both
   - Math Communication study shows that students who have high scores on Content, Organization & Style (COS) tend to do better on open-ended mathematics items than students with low scores on COS – especially when math items require mathematical responses that are *not* writing tasks (e.g., ranking of numbers, drawing graphs, drawing geometric figures, and other numeric and graphic representations).

[5] Separate reports available

ADDITIONAL CONSTRUCT VALIDITY STUDIES

7. Study to examine whether items function differentially for girls vs. boys or whites vs. non-whites (being conducted by UW):

   Results to date suggest that girls and minorities tend to earn their scores from open-ended items; boys and whites from multiple-choice items (which may suggest that tests composed exclusively of multiple-choice items are biased in favor of whites and boys)

8. Study to examine whether WASL reading and mathematics scores predict college freshman GPA (being conducted by UW – including data from UW, WSU, WWU, EWU, CWU):

   Results to date indicate that WASL scores can predict freshman GPA as well or better than SAT scores

RELIABILITY OF SCORES

Reliability refers to whether we can trust (depend up) the scores we get for students or whether there is the possibility of error in the scores.

1. Two critical issues in reliability: error in individual student scores and error in group scores
2. Error is assumed to be randomly positive or negative
3. Causes of error in assessment can be rater inconsistency, student carelessness, leaving items blank, having a bad day, copying others' work, guessing, and other *random* events.
4. Since error is randomly positive or negative, error for groups is smaller than error for individuals because **+**'s and **–**'s cancel each other out
5. Classical test theory measures of score error are estimates of the *average error across all students*
6. Item response theory measures of score error (psychometrics used in WASL) are identified for *each scale score point*.
7. Standard error of the mean is the estimate of error in group scores.
8. Standard error of measurement is the estimate of error in individual scores

POSSIBLE SOURCES OF EVIDENCE FOR THE RELIABILITY OF SCORES

Ways that test developers get *evidence for reliability* of test *scores* for the *individual student*:

1. Rater agreement at the item level. Check to make certain that raters are interpreting scoring rules exactly the same way:
    - Use 'validity papers' – papers that already have scores on them – to see of raters are drifting from the scoring rules
    - Use 'back reads' – have an expert scorer randomly rescore papers
    - Randomly re-score 5-10% of all student work to check for overall rater consistency
2. Rater agreement at the total score level. Check to see if students will get the same total score regardless of the rater.
    - Rater error is likely to be random (if raters are well trained).
    - Raters will sometimes give students higher scores than they should get and sometimes give students lower scores than they should get
    - If error is minor and random, total scores for students should be about the same.
3. Internal consistency. Statistically examine students' responses to items to see if students respond consistently across items within a particular test
4. Test-retest. Give students two parallel test forms OR the same test at two different times. Correlate the students' two scores. High correlations suggest students would get the same score regardless of the test taken.

EVIDENCE FOR THE RELIABILITY OF SCORES

Rater Agreement – Part 1: Check to make certain that raters are interpreting scoring rules exactly the same way

Rater Agreement Grade 10

Year	Content Area	Percent Exact Agreement	Percent Adjacent + Exact Agreement
1999	Reading/Listening	74-97%	99-100%
	Mathematics	70-91%	96-100%
	Writing	84-86%	≈100%
2000	Reading/Listening	80-97%	98-100%
	Mathematics	90-99%	99-100%
	Writing	82-83%	≈100%
2001[6]	Reading/Listening	79-95%	96-100%
	Mathematics	78-98%	98-100%
	Writing	60-71%	95-98%
2002	Writing	66-72%[7]	99%

[6] Change in how rater agreement was computed: Prior to 2001, rater agreement was computed *including* all safeguards to monitor consistency of raters (random "read behinds" by scoring table leaders, random insertion of validity (pre-scored) papers, retraining of raters who drift from scoring rubrics). Beginning in 2001, rater agreement was computed without taking into account use of safeguards. Therefore, percent exact agreement in 2001 and 2002 underestimates rater agreement that influenced students' scores.

[7] First year of teacher involvement in scoring

EVIDENCE FOR THE RELIABILITY OF SCORES

Rater Agreement – Part 2: Check to see if students will get the same total score regardless of the rater.

NOTE: In evaluating the numbers in the tables, keep the following in mind:

- Correlations can range from –1.00 to 1.00.
- A correlation of –1.00 indicates that total scores from two readers would be exactly the opposite.
- A correlation of 1.00 indicates the total scores students would earn from two scorers would be the same or very nearly the same.
- The closer the correlation between total scores from different readers is to 1.00, the better the reliability
- The more similar the means (average scores across all the students), the more likely that the students' first and second total scores were the same or nearly the same

EVIDENCE FOR THE RELIABILITY OF SCORES

1999 Grade 10 Correlations between and Means of Total Scores of First and Second Readings for Open-Ended Items by Test

	Correlation	Mean of Scores from First Reading	Mean Scores of from Second reading
Listening/Reading	.97	16.24	16.03
Writing	.96	6.99	6.95
Mathematics	.98	16.85	16.85

2000 Grade 10 Correlations between and Means of Total Scores of First and Second Readings for Open-Ended Items by Test

	Correlation	Mean of Scores from First Reading	Mean Scores of from Second reading
Listening/Reading	.99	18.22	18.06
Writing	.95	6.38	6.38
Mathematics	.99	15.41	15.39

2001[8] Grade 10 Correlations between and Means of Total Scores of First and Second Readings for Open-Ended Items by Test

	Correlation	Mean of Scores from First Reading	Mean Scores of from Second reading
Listening/Reading	.99	15.15	15.07
Writing	.95	6.73	6.45
Mathematics	.99	11.26	11.25

[8] change in how rater agreement was calculated (see previous page)

EVIDENCE FOR THE RELIABILITY OF SCORES

Internal consistency: Statistically examine students' responses to items to see if students respond consistently across items within a particular test (Classical Test Theory method for estimating error)

1999 Grade 10 Reliability Estimates and Standard Error of Measurement for Scores on each WASL Test

Subtest	Alpha Coefficient	Scaled Score† or Raw Score Standard Error* of Measurement
Listening†	.77	27.6
Reading†	.92	8.4
Mathematics†	.93	11.4
Writing*	.85	1.0

2000 Grade 10 Reliability Estimates and Standard Error of Measurement for Scores on each WASL Test

Subtest	Alpha Coefficient	Scaled Score† or Raw Score Standard Error* of Measurement
Listening†	.62	33.14
Reading†	.90	9.55
Mathematics†	.92	11.3
Writing*	.76	1.12

2001 Grade 10 Reliability Estimates and Standard Error of Measurement for Scores on each WASL Test

Subtest	Alpha Coefficient	Scaled Score† or Raw Score Standard Error* of Measurement
Listening†	.77	30.08
Reading†	.90	9.69
Mathematics†	.92	11.60◄
Writing*	.81	1.00

◄ Using Item Response Theory, the standard error at the cut score is 9.45